38–41 Nucleic Acids Research, 2002, Vol. 30, No. 1

The Ensembl genome database project

T. Hubbard, D. Barker, E. Birney[1,*], G. Cameron[1], Y. Chen[1], L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, R. Durbin, E. Eyras, J. Gilbert, M. Hammond[1], L. Huminiecki[1], A. Kasprzyk[1], H. Lehvaslaiho[1], P. Lijnzaad[1], C. Melsopp[1], E. Mongin[1], R. Pettett, M. Pocock, S. Potter, A. Rust[1], E. Schmidt[1], S. Searle, G. Slater[1], J. Smith, W. Spooner, A. Stabenau[1], J. Stalker, E. Stupka[1], A. Ureta-Vidal[1], I. Vastrik[1] and M. Clamp

The Wellcome Trust Sanger Institute and [1]European Bioinformatics Institute (EMBL–EBI), Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK

Received August 20, 2001; Revised and Accepted October 31, 2001

ABSTRACT

The Ensembl (http://www.ensembl.org/) database project provides a bioinformatics framework to organise biology around the sequences of large genomes. It is a comprehensive source of stable automatic annotation of the human genome sequence, with confirmed gene predictions that have been integrated with external data sources, and is available as either an interactive web site or as flat files. It is also an open source software engineering project to develop a portable system able to handle very large genomes and associated requirements from sequence analysis to data storage and visualisation. The Ensembl site is one of the leading sources of human genome sequence annotation and provided much of the analysis for publication by the international human genome project of the draft genome. The Ensembl system is being installed around the world in both companies and academic sites on machines ranging from supercomputers to laptops.

INTRODUCTION

A genome sequence provides a natural framework about which to organise biological data. In the short time in which genome sequences have been available, genome databases have proved invaluable resources to researchers. In the case of human, the range of existing biological data and the types of researchers is even wider than for other organisms, stretching from clinical genetics to molecular biology. The availability of the draft human genome sequence enables these huge amounts of data, ranging from records of disease in our species to the sequences of related organisms, to be brought together systematically for the first time.

The Ensembl project is actively addressing this by providing a database of human genome annotation (http://www.ensembl.org/). This is being continuously expanded to include an increasing range of data types (vertical integration) as well as to build comparative genome sequence views as sequences of vertebrate genomes, such as mouse, rat and zebrafish, become available (horizontal integration). The database is being built on a very general and carefully engineered software framework that is being developed in parallel with the data integration. By making all software freely available and designing the system to be completely portable, Ensembl aims to provide a bioinformatics framework that is easy to apply to different organisms and types of data. The hope is that in the spirit of open source community projects such as Linux, Ensembl will be widely adopted and will allow database researchers and developers more time to focus on innovation.

Ensembl GENOME ANNOTATION

Ensembl annotates known genes and predicts novel genes, with functional annotation from the InterPro (1) protein family databases and with additional annotation by OMIM disease (2), SAGE expression (3,4) and by gene family (5).

Prediction of genes is the most important part of genome annotation, connecting the DNA sequence with the wide array of experimental data. In eukaryotic organisms with large introns, *ab initio* predictions are useful but have a high false positive rate and often predict partially incorrect gene structures. Thus, incorporation of all available evidence for gene prediction is necessary.

The Ensembl gene build system incorporates a wide range of methods including *ab initio* gene predictions, homology and gene prediction HMMs. Genes are placed in the genome using a three step process. First, 'best in genome' positions for all known human proteins from SPTREMBL (6) are found using a fast protein to DNA matcher (pmatch, R. Durbin, unpublished software). These positions are refined using genewise (7) to provide an accurate gene structure. UTRs are also aligned to each gene structure using full-length cDNAs where known. Secondly, a similar process is used to align paralogous human proteins and proteins from other organisms to the genome to form a set of novel human genes. Finally, the *ab initio* program genscan (8) is run across the entire genome to create a set of genscan peptides. Exons from these predicted peptides that are confirmed by blast matches to proteins, vertebrate mRNA and UniGene clusters are assembled into genes.

The above process creates a set of transcripts and these are grouped into genes wherever an exon is shared. These 'Ensembl genes' are regarded as being accurate predicted gene structures with a low false positive rate, since they are all supported by experimental evidence of at least one form via sequence

*To whom correspondence should be addressed. Tel: +44 1223 494420; Fax: +44 1223 494468; Email: birney@ebi.ac.uk

homology. Ensembl human genes are identified by numbers beginning ENSG (transcripts begin ENST, exons begin ENSE and translations begin ENSP). These identifiers are keep stable, as far as is possible, between assemblies of the human genome.

Ensembl is continuously refining and extending its gene building process, calibrating it against regions of the genome that have been hand annotated and experimentally investigated, such as human chromosome 22 (9). We are in the process of integrating EST data into Ensembl gene building. ESTs offer a considerable advantage in aiding the prediction of non-coding exons, especially those located within the 3′-UTR. Two EST/genome alignment algorithms, namely exonerate (G. Slater, unpublished) and EST_genome (10), have been integrated with the Ensembl gene-building pipeline to yield gene predictions incorporating EST alignments. Because EST data are notorious for their high sequence error rate, strict quality measures have been introduced such that only splicing ESTs are considered, and priority is given to those ESTs which align on the genome into clusters.

The whole genome shotgun (WGS) sequence of the mouse genome (data generated by the mouse sequencing consortium) is another rich source for identifying human genes. We have developed a very fast gapped DNA–DNA alignment algorithm 'exonerate' and have used it to align 14 million mouse reads to the assembled human genome. We have found that matches between human and mouse can be assessed using genscan to indicate those which are potentially novel coding exons.

Ensembl WEB SITE

The Ensembl automatic annotation of the human genome sequence is available as an interactive web service (http://www.ensembl.org/).

A view of a region of genomic sequence is shown in Figure 1. Ensembl *contigview* web pages feature the ability to scroll along entire chromosomes, while viewing the features within a selected region in detail. Features are integrated from external data sources such as HUGO gene names, genetic markers, disease genes and SNPs, with links to primary databases. The user can control which features are displayed and dynamically integrate external DAS data sources as well as their own annotation (see below). Matches between WGS mouse reads and the human genome from exonerate are also displayed. The individual mouse reads can be accessed via the EBI trace server which is also provided via Ensembl (http://trace.ensembl.org/). There is an integrated, context sensitive, searchable help system which can be accessed by selecting the help button on any page.

The Ensembl web site provides a variety of alternate views of the data. These include *mapview* web pages, which show relationships between cytogenetic bands and the genome sequence via markers, and displays feature distribution plots; *geneview* web pages showing information about individual Ensembl gene with its transcripts and gene structures and *proteinview* web pages, showing information about individual Ensembl translations with functional annotation from InterPro. Similarity searching is also integrated into the web site. BLAST (11) and SSAHA (http://www.sanger.ac.uk/Software/analysis/SSAHA/) search tools are available against the entire human genome sequence, predicted gene datasets and mouse genome trace and whole genome assembly datasets.

Ensembl can be accessed in a variety of ways apart from web pages. Ensembl annotation can also be viewed interactively using the Apollo Java viewer, which is being developed as a collaborative project between the Berkeley *Drosophila* Genome Project (http://www.bdgp.org/) and Ensembl. The Ensembl FTP site provides a variety of data download formats, e.g. FASTA files of gene and protein sequences; EMBL and GenBank formats containing annotation of the raw genomic sequence. This includes the full dumps of the MySQL database used by the web site (see below). Extensive data dumping tools are also available from the *contigview* web pages, allowing regions to be selected and dumped in many flat file formats. Regions can also be dumped as graphical images for printing in a variety of formats and layouts.

Currently Ensembl has annotated human and mouse sequence available via its web site. We are in the process of annotating worm, fly, fugu and mosquito in collaboration with their respective genome communities.

Ensembl SOFTWARE SYSTEM

To achieve scalability and consistency of annotation we have developed a portable software system based around a relational database and a series of reusable components. We use Bioperl as a base bioinformatics library (http://www.bioperl.org/) and the free MySQL relational database. The entire Ensembl source code is freely available under an Apache open source licence. It is mainly written in Perl, but with extensions in C and some alternative interfaces are available in Java.

The architecture of the software is split into biologically meaningful objects (business objects) and database connectivity objects (adaptors). This separation makes it easier to evolve the schema to address new datatypes or analyses while maintaining code stability. New datasets can be added easily by providing the necessary adaptors and business objects.

One of the core design features of the system is the 'Virtual Contig' (VC) object, which allows access to genomic sequence and its annotation as if it was a continous piece of DNA in a 1–N coordinate space, regardless of how it is stored in the database. This is important since it is impractical to store large genome sequences as continuous pieces of DNA, not least because this would mean updating the entire genome entry whenever any single base changed. The VC object handles reading and writing of features and behaves identically regardless of whether the underlying sequence is stored as a single real piece of DNA (a single raw contig) or an assembly of many fragments of DNA (many raw contigs). Because features are always stored at the raw contig level, 'virtual contigs' really are virtual and as a result are less fragile to sequence assembly changes. It is this feature that allows Ensembl to handle draft genome data in a seamless way and makes it possible to change between different genome assemblies relatively painlessly. This feature should also put us in a good position to handle haplotype sequences efficiently as they become available.

Access to the software is via FTP to stable snapshots or via a CVS server to live development code. As an open source project we have an active community of both academic and commercial developers using CVS. The entire system is portable as well as its individual components, including the web site and analysis pipeline. This allows users to install the system to process their own genome data as well. By downloading the MySQL

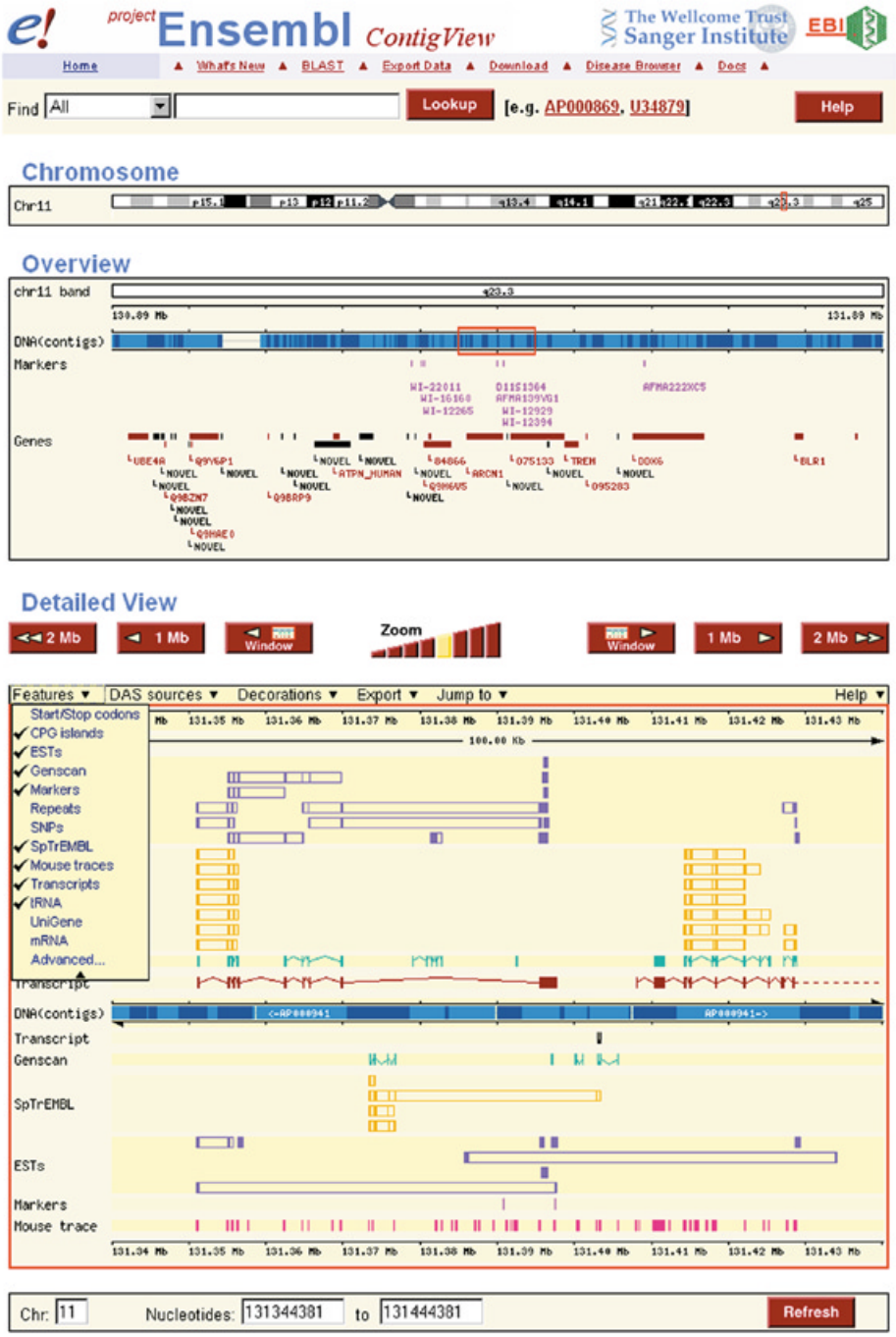

Figure 1. Screenshot of Ensembl contigview, showing the region of human chromosome 11 around genome sequence accession AP000869. The region is shown at three resolutions and navigation (re-centre on click) is possible by clicking in any of the three panels. In the top 'Chromosome' panel a red box shows the region being viewed in q23.3 with respect to the cytogenetic banding pattern of the entire chromosome. In the middle 'Overview' panel a second red box similarly shows the region being viewed in detail below. The middle panel shows the location of markers and genes, by default >1 Mb. Genes are coloured brown and labelled with either HUGO identifiers or SPTREMBL IDs if they are known. Novel 'Ensembl genes' (see text for definition) are labelled as such and shown in black. Annotated genes from EMBL/GenBank sequence files, where present, are shown in green. The lower 'Detailed View' panel shows genomic sequence features in detail, by default >100 Kb. Gene colour scheme is as for the 'Overview' panel, with the addition of sequence contig based genscan *ab initio* predictions shown in cyan. Matches to SPTREMBL entries are shown in yellow, with boxes linking a series of matches to the same entry. Matches to the WGS mouse genome are shown in purple. The region being viewed can be zoomed and re-centred with the mouse or specified precisely in chromosomal coordinates. The pull-down menu shown is one of several and allows the user to select the features being displayed. The second pull down allows the addition of annotation from third party DAS sources. Floating menus (not shown) appear as the mouse is moved over any feature, allowing access to pages with additional information.

dumps it is also possible to setup a full mirror of the pre-computed analysis of the human genome provided by the Ensembl web site. Currently there are over 20 remote installations of the web site. However, the power of the system is not limited to a web interface: the object interfaces, such as 'virtual contigs', to our pre-computed data stored in MySQL provide a new way for

research groups to carry out analysis of human genome data without the huge effort of having to first organise the raw sequence.

Ensembl DATA ANALYSIS PIPELINE

The human genome sequence is more than an order of magnitude larger than the previous largest genomes of worm and fly, which are in themselves an order of magnitude larger than most of the other genomes that have been sequenced. Also, the human genome sequence is made up of fragments and is rapidly changing as the draft sequence is finished (now >50% of the genome is finished). Ensembl works closely with the primary providers of data in the international human genome sequencing consortium (12) and hosts one of the two 'Genome Central' sites, with links to primary HGP data sources (http://www.ensembl.org/genome/central/). Ensembl currently tracks the sequence assemblies (refered to as the 'golden path') provided by Jim Kent at UCSC and the Ensembl web interface links at the DNA level to the UCSC web interface (http://genome.cse.ucsc.edu). Ensembl reanalyses the human genome whenever a new assembly become available, maintaining the stability of its gene identifiers between releases wherever possible (see above).

Being able to handle the required scale of analysis, which is dynamic as a result of a continuously changing assembly as opposed to static for the storage and display of static data, has been one of the major challenges for the Ensembl project. For example, it requires many millions of individual BLAST sequence comparisons alone to be run successfully without any failures. To make this possible, the Ensembl software system contains a full analysis pipeline. The analysis components are designed around two generic interfaces, one of which encapsulates running a single analysis process and another which encapsulates reading and writing the input and results of an analysis from a database. This separation allows us to write new components rapidly and in particular allows the construction of composite processes. These generic interfaces are then controlled by a scheduling system, which can handle dependencies and retries on top of low level task schedulers, such as LSF.

DISTRIBUTED ANNOTATION SYSTEM (DAS)

While Ensembl aims to provide baseline annotation, genomes are far too complex for any organisation to have a monopoly of ideas or data (13). Ensembl has been actively developing software to support the DAS standard (http://www.biodas.org/) (14), to enable users to easily view and compare annotation from different sources that are distributed across the Internet. Traditionally, different sources of information have been integrated on the Internet via links. However, from the user's point of view this means continously jumping from one data provider's user interface to another and also makes it very difficult to compare, for example, several alternative gene predictions. DAS addresses this through clients which integrate data served by from a number of different DAS servers.

Ensembl makes use of DAS in several ways. First, it makes its annotation data available (http://servlet.sanger.ac.uk:8080/das/) using the biojava DAS server DAZZLE (http://www.biojava.org/) for users with third party DAS clients. Secondly, for users who want to view annotation from human genome DAS servers without setting up third party clients, Ensembl *contigview* can be configured to act as a DAS client (by default *contigview* is pre-configured with a selection of useful DAS sources). Thirdly, for users who wish to serve DAS without setting up a server, limited amounts of user annotation can be uploaded to the Ensembl DAS server.

CONTACTING Ensembl

Ensembl is a joint project of the European Bioinformatics Institute (EBI) and the Sanger Centre, both of which are located on the Wellcome Trust Genome Campus, Cambridge, UK. To receive announcements about updates, subscribe to the 'announce' mailing list: majordomo@ebi.ac.uk 'subscribe ensembl-announce'. To follow the day to day development of Ensembl subscribe to the 'development' mailing list: majordomo@ebi.ac.uk 'subscribe ensembl-dev'. Requests for information and support can be sent to helpdesk@ensembl.org, which is a fully supported helpdesk. Extensive additional documentation can be found on the Ensembl web site, including installation guides and tutorials, both about using the software system and the web interface.

ACKNOWLEDGEMENTS

We are grateful to users of our web site and the developers on our mailing lists for much useful feedback and discussion. The Ensembl project is principally funded by the Wellcome Trust with additional funding from EMBL.

REFERENCES

1. Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Birney,E., Biswas,M., Bucher,P., Cerutti,L., Corpet,F., Croning,M.D. *et al.* (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, **29**, 37–40.
2. Antonarakis,S.E. and McKusick,V.A. (2000) OMIM passes the 1,000-disease-gene mark. *Nature Genet.*, **25**, 11.
3. Velculescu,V.E., Zhang,L., Vogelstein,B. and Kinzler,K.W. (1995) Serial analysis of gene expression. *Science*, **270**, 484–487.
4. Wheeler,D.L., Church,D.M., Lash,A.E., Leipe,D.D., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Tatusova,T.A., Wagner,L. and Rapp,B.A. (2001) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **29**, 11–16. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 13–16.
5. Enright,A.J., Iliopoulos,I., Kyrpides,N.C. and Ouzounis,C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.
6. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
7. Birney,E. and Durbin,R. (2000) Using GeneWise in the *Drosophila* annotation experiment. *Genome Res.*, **10**, 547–548.
8. Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.
9. Dunham,I., Shimizu,N., Roe,B.A., Chissoe,S., Hunt,A.R., Collins,J.E., Bruskiewich,R., Beare,D.M., Clamp,M., Smink,L.J. *et al.* (1999) The DNA sequence of human chromosome 22. *Nature*, **402**, 489–495.
10. Mott,R. (1997) EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. *Comp. Appl. Biosci.*, **13**, 477–478.
11. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J.H., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
12. International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
13. Hubbard,T.J.P. and Birney,E. (2000) Open annotation offers a democratic solution to genome sequencing. *Nature*, **403**, 825.
14. Dowell,R.D., Jokerst,R.M., Day,A., Eddy,S.R. and Stein,L. (2001) The Distributed Annotation System. *BMC Bioinformatics*, **2**, 7.